Indonesian Journal of Electrical Engineering and Computer Science
Vol. 17, No. 3, March 2020, pp. 1524~1530
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v17.i3.pp1524-1530

Medical documents classification using topic modeling

Maryam Nuser[1], Enas Al-Horani[2]
[1,2]Computer Information Department, Faculty of Information Technology and Computer Science, Yarmouk University, Jordan
[1]Computer Science Department, Faculty of Computer & Information Technology, Jordan University of Science and Technology, Jordan

Article Info

Article history:

Received Mar 14, 2019
Revised Jul 30, 2019
Accepted Oct 21, 2019

Keywords:

Classification
Latent dirichlet allocation
Medical documents
Mining health data
Topic modeling

ABSTRACT

The number of digital medical documents is increasing continuously; several medical websites share a lot of unclassified articles. These articles have very long texts that should be read to determine the topic of each document. The classification of these documents is important so researchers can use these documents easily and the effort and time in reading and searching for a specific topic will be reduced. Therefore, an automatic way to extract latent topics from these text documents is needed. Topic modeling is one of the techniques used to deal with this problem. In this paper, a medical collection of documents is used; this collection contains documents from three types of widespread diseases (Heart Diseases, Blood Pressure and Cholesterol). LDA topic modeling technique is applied to classify these documents into the previous mentioned topics. An evaluation of the algorithm's results is done and the LDA shows a good level of classification accuracy.

Corresponding Author:

Maryam Nuser,
Computer Information Department,
Yarmouk University, 21163, Irbid, Jordan.
Email: mnuser@yu.edu.jo

1. INTRODUCTION

The health sector in any country should emphasize on having a healthy community. Advancement in health care is based on previous research in the field. Researchers need to search, read, analyze, and explore published documents in order to follow up with the progress researchers made in the field. Nowadays the number of electronic documents archived is increasing, and becoming harder to organize and understand, so to deal with this large number of documents a need arises to some techniques or computational tools to automatically organize these collections of documents. In addition, efficient search and browse should be considered.

Existing search techniques try to match words in the query with the words in the documents to return documents that contain the questioned words. Words have multiple meanings, and therefore matching between words in the query and documents is not enough to retrieve the documents that are compatible with the user's conceptual topic or meaning. Therefore, words in the same sentence should be considered rather than words separately. Researchers of machine learning and statistics used hierarchical probabilistic models called topic models to build new methods to find patterns of words from a collection of documents. These patterns reveal the topics contained in the documents. These hierarchical probabilistic models can be used with various kinds of data that ranges from words, images, and to survey information [1, 2]

A Topic model is one type of statistical models that is used to discover the abstract topics of the document collection and it can also be thought of as a form of text mining, to obtain patterns of words in textual material. There are various kinds of topic models such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM). LDA is the one that will be used in this research.

1.1. Latent Semantic analysis (LSA)

LSA is a natural language processing technique that investigates the relationships between a set of documents and the terms they contain based on the distributional hypothesis. A vector space is created that contains words counts per paragraph for each text document. Singular value decomposition (SVD) technique is then applied. Words are then compared to decide on the similarity between words. LSA helps in finding information beyond the lexical level of word occurrences; it provides semantic relations between words and documents [3, 4].

A method to handle observed term-document association data statistically was proposed in [5], they assumed that there is underlying latent semantic structure in the data with randomness of word choice with respect to retrieval. They applied Latent Semantic Analysis (LSA) in order to estimate this latent structure and the noise of words. They created a semantic space for a large matrix of term-document association data in which terms and documents that are closely associated are placed next to each other.

1.2. Probabilistic Latent Semantic Analysis (PLSA)

PLSA is derived from a statistical view of LSA. It defines a generative data model that can be used in information retrieval, machine learning, natural language processing, and in related areas. PLSA is proposed to deal with the weaknesses of LSA that uses Singular Value Decomposition of co-occurrence tables; PLSA is based on a mixture decomposition derived from a latent class model it associates a latent context variable with each occurrence of word, which takes polysemy into consideration. There are two main advantages of PLSA: 1) Perplexity minimization for a document-specific unigram baseline. 2) Automated indexing of documents. One way to compare predictive performance of PLSA and LSA is to specify how to extract probabilities from LSA decomposition. The PLSA outperforms the LSA in perplexity reduction relating to the unigram baseline and shows improvements over Latent Semantic Analysis in a number of experiments [6, 7].

1.3. Latent Dirichlet Allocation (LDA)

LDA is a generative statistical model for collection of text data. LDA is a three level hierarchical Bayesian model; each document of a collection is modeled as a mixture of various topics. Each topic is modeled as a mixture over a set of topic probabilities. In the text modeling, each topic probabilities provide an explicit representation of a document [8]. LDA deals with the words of the documents as a bag of words (it means that the order of the words in the document is not considered). The document is represented by term-document matrix that contains the occurrences of each word in each document of the collection [8, 1].

The main idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA proposed the following generative process [8] for each document w in a corpus D:

1) Choose the number of words N according to Poisson distribution.
2) Then choose a topic mixture for the document according to Dirichlet distribution. Dir(α).

A high value of α means that every document is likely to contain a mixture of most of the topics not just a single topic, low value of α means that a document is more likely to be presented by mixture of a few of the topics, so high α makes documents more similar to each other.

3) for each of the N words:
   a) Choose a topic zn according to Multinomial distribution.
   b) generate a word (wn) according to multinomial probability conditioned on the topic (zn).

A document is a probability distribution over topics. A topic is a probability distribution over words. Words that appear in the same document are related. The model generates a document by taking the right number of words from specified topic and mixing them together. Every document is a collection of words that are taken from different topics.

The model try to produce topic distribution, the distribution will have as many topics as we asked the model to make and the highest value of probabilities of words distribution present the fraction of words in the document that originated from a given topic.

The result of LDA is a file that contains all topics made of the words with probabilities belonging to the topic. (Each document represented as a pattern of LDA topics).

1.4. Research Motivation

Several researches were conducted on health information systems and medical data[9-15]. Some researchers worked on the classification of different diseases such as diabetics [11], Alzheimer [12], cancer [13, 14] while others compared several classification and data mining algorithms on health data [15] whether these data were in English, Arabic[16], or multilingual[17, 18]. There are many applications on topic modeling that were applied in different domains by different topic modeling approaches. The literature

contains many examples on researchers who used topic modelling and especially LDA for either text classification [19, 20] or medical diagnosis [21, 22] and the method showed its efficiency[23, 24].

As a reason of unclassified documents and the difficulty to read and determine the topic of each document in the medical document collection, the classification should be done automatically by using topic models to make it possible to obtain the needed documents in a specific topic. The main objective of this research is to use LDA method on a collection of medical documents to classify these documents over three main topics that are strongly related to each other.

2. RESEARCH METHOD

This research is done in a series of operations to classify the collection of documents by using LDA topic modeling and study the performance of this technique on these documents, which can be summarized by Figure 1 as follows:

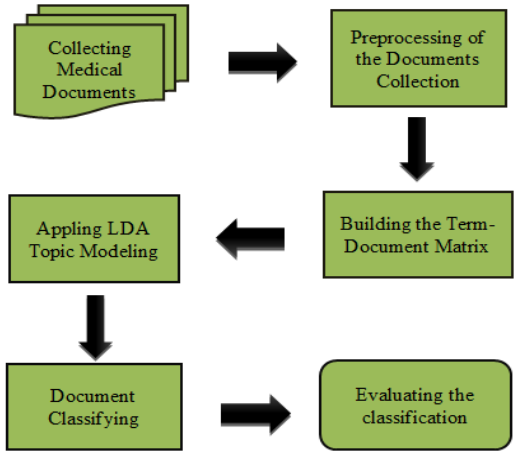

Figure 1. The methodology main steps

The phases followed in this research for classifying medical documents are as follows:

2.1. Data Collection:

Medical articles have very long texts and contain many sections such as the abstract, introduction, materials and other sections about the diseases and their treatments….etc. In addition, there are a lot of tests with numbers and measurements that need to be recorded.

The collection that is used in this research is gathered from medical web sites. The data set contains 500 documents of medical articles that are collected from three medical websites: Medscape (http://www.medscape.com), Hindawi website (http://www.hindawi.com) and PubMed (http://www.cbi.nlm.nih.gov/pubmed/). These websites provide free access to many articles. Each document contains the abstract, conclusion and keywords of each article. The reason of choosing abstract, conclusion and keywords section of the article is that these parts represent the idea or summarize each article and contain the important words in the subject of the article.

The collected medical documents are chosen from three categories: Heart Diseases, Blood Pressure or Hypertension and Cholesterol or Hyperlipidemia. 165 documents are about Heart Diseases, 181 documents are about Blood Pressure and 154 are about Cholesterol.

2.2. Preprocessing and Cleaning:

Preprocessing and cleaning the documents from irrelevant data is an important step for any model [25] which will improve the results. This is the most important step in text analysis. Unclean data has a negative effect on the results. In this step, the collected documents from the previous step (that are saved in Notepad files) are cleaned, and the necessary preprocessing is done in order to make the documents ready to use.

To carry out the implementation for preprocessing and topic modeling; R tool (R Studio) will be used. R Studio language tool is one of the most powerful and popular free software environments. R is a

language and environment for statistical computing and graphics. Furthermore, R provides a wide variety of statistics such as T-test, classification and clustering. Furthermost, R applies graphical techniques, and is highly extensible.

The tm package (in R Studio tool) offers number of transactions that ease the process of cleaning data. For example, the corpus is cleaned using a known cleaning process such as removing Numbers, removing Punctuations, removing special Characters (@, #, %..), removing strip whitespaces, removing Stopwords (include English stop words like or, and, the…etc). In addition, the tool allows the user to add a list of words to the stop words list. As for example, the words from the collected data that have the lowest weight such as Abstract, Conclusion, Keywords, can be added [26].

The common English stop words from http://www.ranks.nl/stopwords website are used here. The 500 documents were input to the tm package for preprocessing and cleaning. A corpus of 158 different terms was resulted and will be used to classify the documents.

2.3. Building the Document Term Matrix

At this step, the Document-Term Matrix (DTM) is created, a matrix that lists all occurrences of words for each document in the corpus. In the DTM the rows represent the documents (each row labeled or start with document's name) and the columns represent the terms (or words) of the documents, beside each document (or row) there are numbers 0, 1, 2, 3…n as entry under each column (term), this number means how many times a term occurs in specific document, if the matrix entry of one row (ex. Doc1) and under one of the columns (ex.term1) is zero, it means that this term doesn't occur in this document otherwise it is possible to be 1, 2, 3…n where n is the frequency of that term.

A list of terms of the matrix with their frequencies is sorted by their frequencies. Words with low frequencies were removed from the corpus in order to reduce the sparsity of the matrix. The sparsity was reduced from 99% to 82%. Furthermore, words that occur with high frequency in the corpus and are not important for the classification process, such as "abstract", "keyword", are also removed. 158 terms from the corpus of 500 documents will be used to classify the corpus into the suggested topics.

2.4. Applying Topic Modeling on the Medical Documents

The DTM that resulted from the previous step is used as an input to this phase. To apply LDA topic modeling, the topic modeling package is used and the number of topics is specified as 3 because the documents in the corpus were chosen from three different medical subjects (we choose 3 topics because we need to classify the documents to their real subjects from the three diseases).

The output will be three topics (Heart Disease (Topic 1), Blood Pressure (Topic 2) and Cholesterol (Topic 3)) each one with associated terms that are related to that topic with different probabilities. Table 1 shows the top 10 terms associated with each topic. It shows that for example, documents in topic 1 have a high probability of containing the words hypertension, study, blood,… while documents in topic 2 are more probable to have the words risk, disease, cardiovascular and so on.

It should be emphasized that each document is considered to be a mixture of all topics (three topics in this research) and each topic contains all terms in the corpus with different probabilities. Table 2 shows the assigned probabilities of the first 12 documents to the three topics. The table indicates that document number 1 is classified as topic 1 with a probability of 0.31, at the same time it is classified as topic 2 with a 0.40 probability; and it is classified as topic 3 with a 0.28 probability.

Table 1. The top 10 Terms Related Each Topic

Topic 1	Topic 2	Topic 3
hypertens	Risk	cholesterol
Studi	Diseas	level
Blood	cardiovascular	patient
pressur	Heart	effect
patient	Clinic	treatment
Age	Chd	increas
control	Patient	therapi
signific	Coronary	reduc
Group	Outcome	Lipid
higher	Medic	lower

Table 2. Assigning Documents Probabilities to the Topics

Topic1	Topic2	Topic3
0.313131	0.40404	0.282828
0.295833	0.433333	0.270833
0.220238	0.264881	0.514881
0.276423	0.227642	0.495935
0.422572	0.186352	0.391076
0.307359	0.307359	0.385281
0.278867	0.206972	0.514161
0.280303	0.405303	0.314394
0.185535	0.279874	0.534591
0.290196	0.254902	0.454902
0.22807	0.259649	0.512281
0.435897	0.238095	0.326007

2.5. Medical Documents Classification

Finally after extracting the topics terms, each document has three values of assigned probabilities to the three topics. The Topic with the highest probability will be chosen to classify the document. As a result, Document number 1 is classified as topic 2, Document number 2 is also classified as topic 2, while document number 3 is classified as topic 3. These results are highlighted in Table 3 for the first 12 documents of the corpus.

Table 3. List of the Documents Topic Assignments

Document	Topic1	Topic2	Topic3	Assigned Topic
1.txt	0.313131	**0.40404**	0.282828	2
10.txt	0.295833	**0.433333**	0.270833	2
100.txt	0.220238	0.264881	**0.514881**	3
101.txt	0.276423	0.227642	**0.495935**	3
102.txt	**0.422572**	0.186352	0.391076	1
103.txt	0.307359	0.307359	**0.385281**	3
104.txt	0.278867	0.206972	**0.514161**	3
105.txt	0.280303	**0.405303**	0.314394	2
106.txt	0.185535	0.279874	**0.534591**	3
107.txt	0.290196	0.254902	**0.454902**	3
108.txt	0.22807	0.259649	**0.512281**	3
109.txt	**0.435897**	0.238095	0.326007	1

3. RESULTS AND ANALYSIS

3.1. Evaluation of the Accuracy of LDA

Before the evaluation phase starts, the documents were sent to a medical expert in a summarized form. The expert classified the documents as belonging to topic1, 2 or 3. These classifications were used as a base to evaluate the classifications produced from the LDA algorithm.

In this evaluation phase, the LDA classification results will be compared with the classification of the documents that is classified by experts in medical domain to determine the accuracy value of applying this technique on the data set. This value represents the effectiveness of LDA topic modeling in classifying medical text documents. Table 4 shows a sample of the comparison of the results predicted by LDA and classified by the expert for the first 12 documents.

The accuracy is measured by using the Confusion Matrix as shown in Table 5, which shows how many documents are classified correctly and how many documents are misclassified. As for example, from the 181 documents in topic 1, 112 only were classified correctly; while 38 documents were misclassified as topic 2 and 27 documents were misclassified as topic 3.

Table 4. A Comparison between the LDA Documents Classification with the Real Topic of the Documents

Document	predicted topic	Actual topic	Match?
1.txt	2	2	Yes
10.txt	2	2	Yes
100.txt	3	3	Yes
101.txt	3	3	Yes
102.txt	1	3	No
103.txt	3	3	Yes
104.txt	3	3	Yes
105.txt	2	3	No
106.txt	3	3	Yes
107.txt	3	3	Yes
108.txt	3	3	Yes
109.txt	1	1	Yes

Table 5. The Confusion Matrix

Actual Topic	predicted Topic			Total	Accuracy
	Topic1	Topic2	Topic3		
Topic1	**112**	38	27	181	**61.8%**
Topic2	17	**134**	15	165	**81.2%**
Topic3	17	29	**111**	154	**72.1%**
Total	146	201	153	500	
Accuracy	**76.71%**	**66.7%**	**72.5%**		**71.4%**

$$\text{Accuracy} = \frac{\#\text{of correctly classified documents}}{\text{Total number of docuents}}$$

$$= \frac{112+134+111}{500} = 357/500 = 71.4\%$$

3.2. Results Analysis

After applying the LDA model to classify 500 documents into three topics: Topic1 about Blood Pressure or Hypertension, Topic2 about Heart Diseases or Cardiovascular and Topic3 about Cholesterol or Hyperlipidemia, the output was 146 documents were assigned as Topic1, 201 documents assigned as Topic2 and 153 documents assigned as Topic3. The overall Accuracy of the documents classification was 71.4%.

The LDA assigns a probability to each document. All documents were medical documents and this means that they have several words in common. In addition, the chosen diseases (topics) are related to each other. As a result, there is a probability that documents will be misclassified. In addition, in rare cases the LDA assigns probabilities that are close to each other. As for example, document 5 is assigned as topic 1 with probability 42% and as topic 3 with probability 39%. As mentioned before this is because documents have words in common and are all in the same main category (medical documents). Additionally, the preprocessing step is important in affecting the process of extracting the topics.

4. CONCLUSION

Due to the large number of digital medical documents that are not classified into specific subjects or topics and because of the long text in each document and the several sections it has, a need to the classification arises. An automated classification method will reduce the time and effort needed to classification compared to manual classification by a field expert.

One of the most common classification techniques is Topic Modeling. LDA topic model is used in this research to extract topics from the collected documents and assign them to the most probable topic. Five hundred documents were collected from medical websites. Preprocessing is done to the documents, and the results are fed to the LDA tool. The output was 357 documents were correctly classified from the 500 documents in the collection. LDA shows an accuracy of 71.4%.

Studying another Topic Modeling technique like CTM in order to see its performance and comparing it with the results of LDA Model on our collection is a future work that should be considered. A further study on the effect of stopwords removal on the results of the topic model and measure the accuracy of the classification before and after removing them can be done as future work. Another idea is to collect more documents (increasing the size of documents collection) and studying if the size of the collection affect the extraction of topics and the classification of documents or not.

REFERENCES

[1] Alghamdi R and Alfalqi K., "A Survey of Topic Modeling in Text Mining". *International Journal of Advanced Computer Science and Applications*; Vol. 6 No. 1, pp. 147-153, 2015
[2] Blei D M. "Probabilistic Topic Models", *Communications of the ACM*; Vol. 55, No. 4, pp. 77-84,2012.
[3] Landauer T K, Foltz P W and Laham D. "An introduction to latent semantic analysis", *Discourse Processes* Vol. 25, pp. 259-284, 1998.
[4] Dumais S T. "Latent Semantic Analysis", *Annual Review of Information Science and Technology*, Vol. 38, No. 1, PP. 188–230, 2004.
[5] Deerwester S., *et al.*, "Indexing by latent semantic analysis" *Journal of the American society for information science*; Vol. 41, No. 6, pp. 391-407, 1990.
[6] Hofmann T. "Probabilistic latent semantic analysis", *In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* July, pp. 289-296. Morgan Kaufmann Publishers Inc. 1999.
[7] Hofmann T. "Unsupervised learning by probabilistic latent semantic analysis:, *Machine learning*, Vol. 42, No. 1, pp. 177-196, 2001.
[8] Blei D M, *et al.*, "Latent dirichlet allocation", *Journal of machine learning research*; Vol. 3, pp. 993-1022, 2003.
[9] Klaib, A.F. and Nuser, M.S., "Evaluating EHR and Health Care in Jordan According to the International Health Metrics Network (HMN) Framework and Standards: A Case Study of Hakeem", *IEEE Access*, Vol. 7, pp.51457-51465, 2019.
[10] Deepika N, M. and Anand, F.Jerald, "A novel three tier internet of things health monitoring system" *Indonesian Journal of Electrical Engineering and Computer Science* Vol. 15, No. 2, pp. 631-637 2019.
[11] Sinan Adnan and Diwan Alalwan "Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset" *Indonesian Journal of Electrical Engineering and Computer Science* Vol. 14, No. 1, pp.88-95. 2019
[12] Jantana Panyavaraporn and Paramate Horkaew,"Classification of Alzheimer's Disease in PET Scans using MFCC and SVM". *Expert Systems with Applications*, Vol. 5, pp. 1829-1835, 2018.

[13] Mohanad Najm Abdulwahed "Classification of Prostate Cancer using Wavelet Neural Network" *Indonesian Journal of Electrical Engineering and Computer Science* Vol. 12, No. 3, pp. 968-973, 2018,
[14] Mohammed Abdulrazaq Kahya "Classification enhancement of breast cancer histopathological image using penalized logistic regression" *Indonesian Journal of Electrical Engineering and Computer Science* Vol. 13, No. 1, pp. 405-410, 2019.
[15] Ashutosh Kumar Dubey, *et al*., "Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data", Vol. 8, No. 1, pp. 18-29, 2018.
[16] Al-Radaideh Q A and Al-Khateeb S S. "An associative rule-based classifier for Arabic medical text", *International Journal of Knowledge Engineering and Data Mining*, Vol. 3, pp. 255-273, 2015.
[17] Karanikolas N N, *et al*., "*Medical Text Classification based on Text Retrieval techniques*", MEDINF. 1st International Conference on Medical Informatics & Engineering, pp. 375-378. Craiova, Romani, October 9 - 11, 2003
[18] Elberrichi Z, *et al*., "Medical Documents Classification Based on the Domain Ontology MeSH", *International Arab Journal of e-Technology*, Vol. 2, No. 4, pp. 210-215, 2012.
[19] Miha Pavlinek and Vili Podgorelec, "Text classification method based on self-training and LDA topic models ", *Expert Systems With Applications*, Vol. 80, No. 1, pp. 83-93, 2017.
[20] Lubis, F.F., *et al.*, "Topic discovery of online course reviews using LDA with leveraging reviews helpfulness", *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 9, No. 1, pp. 426-438, 2019.
[21] Shamna, P., *et al*., "Content Based Medical Image Retrieval, using Topic and Location Model", *Journal of Biomedical Informatics*, 2019.
[22] Jorge Pérez, *et al.*, "Cardiology record multi-label classification using latent Dirichlet Allocation.", *Computer Methods and Programs in Biomedicine*, Vol. 164, pp. 111–119, 2018.
[23] Nikolenko I S, *et al.*, "Topic modelling for qualitative studies", *Journal of Information Science*, Vol. 43, No. 1, pp. 88–102, 2017.
[24] Rathore A S and Roy D. "Performance of LDA and DCT models", *Journal of Information Science*; Vol. 40, No.3, pp. 281–292, 2014.
[25] Nayak S A, *et al*., "Survey on Pre-Processing Techniques for Text Mining", *International Journal Of Engineering And Computer Science*, Vol. 5, No. 6, pp. 16875-16879, 2016.
[26] Lo R T-W, *et al*., "Automatically building a stopword list for an information retrieval system. *Journal on Digital Information Management*: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR), Vol. 5, pp. 17-24, 2005.

BIOGRAPHIES OF AUTHORS

Maryam S. Nuser is an associate professor at the Computer Information Department, Faculty of Information Technology and Computer Sciences, Yarmouk University, Jordan. She received her BSc degree in Computer Science from Yarmouk University in 1995, Msc degree from the University of Arkansas, USA in 2002, and a PhD degree from the University of Arkansas in 2004 with the same major. She worked as a head of CIS deptartment at Yarmouk University during the period 2006-2008. Dr. Nuser has several publications in local and international journals, conferences, and books

Enas Al-Horani Received a master's degree from the computer Information Systems department, Faculty of Information Technology and Computer Sciences, Yarmouk University, Jordan. She is currently working at extensya company. Her primary areas of interest include Health information systems and data mining.